EXTRACTING THE RELEVANT DELAYS IN TIME SERIES MODELLING

Cyril Goutte
Department of Mathematical Modeling - Bygn. 321
Technical University of Denmark, DK-2800 Lyngby, Denmark
Phone: +45 4525 3921
Fax: +45 4587 2599
E-mail: cg@imm.dtu.dk

Abstract. In this contribution, we suggest a convenient way to use generalisation error to extract the relevant delays from a time-varying process, i.e. the delays that lead to the best prediction performance. We design a generalisation-based algorithm that takes its inspiration from traditional variable selection, and more precisely stepwise forward selection. The method is compared to other forward selection schemes, as well as to a non-parametric tests aimed at estimating the embedding dimension of time series. The final application extends these results to the efficient estimation of FIR filters on some real data.

OVERVIEW

In system identification as well as in time series modelling, the choice of the inputs to our model plays a crucial role. In order to obtain good performance, one shall model future behaviour from a set of *relevant* past measurements. An insufficient amount of inputs will prevent the model from capturing the underlying mapping. On the other hand, including irrelevant inputs will lead to poor prediction performance, as suggested by the "curse of dimensionality".

In this contribution, we consider a method aimed at finding a set of relevant delays. For that purpose, we use a suboptimal iterative method that minimises the estimated generalisation error, and bears resemblance to the usual statistical variable selection methods [6]. However, this Extraction of Relevant Delays (ERD) method is original in the fact that 1) it assesses the relevance of possible inputs on the basis of generalisation, and 2) it is adapted to time dependent problems.

The organisation of this paper is as follows: first we give a short presentation of the topic of statistical variable selection, and describe our ERD method. We then introduce briefly a class of methods estimating the embedding dimension of time series. The second part of the paper contains a number of experiments conducted on the well-known Hénon map, on a real time series, and finally on a FIR filtering problem. We conclude with a discussion of the results.

INPUT SELECTION

Let us consider a standard time series modelling problem. A sequence x of measurements is collected, and we try to predict x_t from a set of past values x_{t-d}. Note that in that setting, the length of the basic time delay (i.e. difference between t and $t+1$) is imposed on us. Extracting the relevant delays consists in finding a set of m delays $(x_{t-d_1}, \ldots, x_{t-d_m})$ that, given as input to a model, yields the best prediction.

This is a special case of variable selection, which in turn can be seen as part of the more general problem of analysing the structure in the data [6]. An important assumption in conventional variable selection is that all necessary variables are available, i.e. a sufficient subset of inputs actually exists. Provided that data are sampled correctly, this assumption is usually satisfied in the case of time series[1]. We will use the terms 'variable', 'input' or 'delay' indifferently when addressing our time series modelling problem.

An exhaustive search through all possible subsets of inputs is usually impossible for combinatorial reasons. A number of suboptimal techniques have thus been designed, among them stepwise methods:

- *Forward selection* methods consists in starting from an empty set of inputs, and adding variables one after the other according to a given *selection criteria*, until a chosen *stopping condition* is fulfilled.
- On the contrary, *backward elimination* methods start with the full set of inputs, and proceed by deleting one variable at a time according to the *selection criteria*, until the *stopping condition* is reached. In the field of neural computation, variable selection techniques based on pruning [2] are a typical example of backward elimination.

Stepwise regression usually refers to a combination of both (in the linear case). For both methods, the crucial parts are the design of the selection criteria, and the stop condition. Conventional methods in linear regression rely on e.g. correlation coefficients, information content or F-testing.

EXTRACTION OF RELEVANT DELAYS

We present here a method of Extraction of Relevant Delays (ERD) that relies upon generalisation error. It draws its inspiration from forward selec-

[1] It breaks down in the case where a long-term delay is needed, that ranges further than the time period spanned by the data. However, the relevance of such long-term prediction is questionable, and there would be no data to identify the associated parameter(s) anyway.

tion methods, combined with generalisation estimation. Consider a model f providing a mapping from an input vector containing m delays $\mathbf{x}^{(t)} = (x_{t-d_i})_{i=1\ldots m}$ to output x_t, and assume Gaussian perturbation on the output. We define the *generalisation error* (or expected risk) for this model as:

$$G(f) = \int \left(f\left(\mathbf{x}^{(t)}\right) - x_t \right)^2 p\left(x_t, \mathbf{x}^{(t)}\right) \, \mathrm{d}x_t \mathrm{d}\mathbf{x}^{(t)} \tag{1}$$

Obviously, equation (1) can not be used directly as the joint input-output probability is unknown. We will thus resort to estimating this error, or rather its average over all possible training sets of a given size N. There are mainly two classes of such estimators: methods such as cross-validation [17] resample the available data, while algebraic estimators [1] rely on statistical arguments.

Consider for example the second option. Many estimators have been proposed in the literature, e.g. Final Prediction Error (FPE) [1], Generalised Prediction Error (GPE) [11], Final Prediction Error for Regularised problems (FPER) [7] or Network Information Criterion (NIC) [12]. We will here settle for an expression similar to GPE, i.e. a FPE where the number of parameters is replaced by the *number of efficient parameters* $\widehat{P}$:

$$\langle \widehat{G} \rangle = \left(\frac{N + \widehat{P}}{N - \widehat{P}} \right) \langle S \rangle \tag{2}$$

where $\langle S \rangle$ is the average training error (over all training sets of size N). As such an average is not available, we plug the measured training error (or empirical risk) $S(f)$ instead. For quadratic risk, $S(f) = \sum \left(f\left(\mathbf{x}^{(t)}\right) - x_t \right)^2$. The calculation of $\widehat{P}$ depends on the regularisation method used during training (see e.g. [7, 3]).

The proposed ERD method is a forward method taking all delays in their natural order (which bypasses the *selection criteria*), and adds a candidate input if and only if it corresponds to a *significant* decrease in generalisation error. The algorithm can be described as follows:

1. Initialise: $d = 0$; $G_{min} = \sigma_x^2$; no input selected.
2. Model: $d = d + 1$; add delay $t - d$ to selected inputs; estimate generalisation error $\widehat{G}$ for resulting model.
3. Test: if $\widehat{G}$ is significantly smaller than G_{min}, keep delay $t - d$; $G_{min} = \widehat{G}$. Discard otherwise.
4. Iterate: Go to step 2 until stop condition is reached.

Significant decrease in error. When a candidate delay yields a decrease in (estimated) generalisation error, step 3 requires that we assess the significance of this decrease. We take advantage of the fact that the generalisation estimators mentioned above are based on averaging a statistics, and test whether the statistics associated with two different generalisation estimators have statistically significantly distinct means by performing a paired t-test [15, 8].

In our case, the estimated (average) generalisation error given by FPE can be expressed as the following average:

$$\langle \widehat{G} \rangle = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{N + \widehat{P}}{N - \widehat{P}} \right) e_w^{(k)} \tag{3}$$

where $e_w^{(k)}$ is the local risk (e.g. squared residuals) for training example k and a model parameterised by w. Let us consider two models trained on the same set of examples, and $\widehat{P}_1$ and $\widehat{P}_2$ the *numbers of efficient parameters* for the first and second model (respectively). The distribution of the corrected residuals $\left(\frac{N+\widehat{P}_1}{N-\widehat{P}_1}\right) e_{w_1}^{(k)}$ (resp. $\left(\frac{N+\widehat{P}_2}{N-\widehat{P}_2}\right) e_{w_2}^{(k)}$) has mean $\langle \widehat{G}_1 \rangle$ (resp. $\langle \widehat{G}_2 \rangle$). We thus test whether $\langle \widehat{G}_2 \rangle$ is significantly smaller than $\langle \widehat{G}_1 \rangle$ by using a paired t-test on the corrected residuals.

The case of cross-validation is somewhat more straightforward. The *leave-one-out* (LOO) cross-validation score is calculated by averaging the prediction error on one example for a model trained on the remaining sample:

$$\langle \widehat{G} \rangle = \frac{1}{N} \sum_{t=1}^{N} \left(f_t \left(\mathbf{x}^{(t)} \right) - x_t \right)^2 \tag{4}$$

Where f_t is the model trained *without* example $\left(\mathbf{x}^{(t)}, x_t\right)$. For two different models, the residuals are paired according to the example left out, so that a (paired) t-test can be used to determine whether these residuals come from distribution with different mean, i.e. correspond to different average generalisation error. Extension to *m-fold* cross-validation is straightforward.

EMBEDDING DIMENSION

In the study of non-linear dynamical systems, and time series in particular, an important problem lies in finding the *embedding dimension* [16], which is essentially equivalent to finding the set of *primary* delays in time series. In the realm of neural computation, the recently proposed δ-test method[14] addresses this issue. In a different field, a method for identifying the order of non-linear input-output systems was proposed [5], that relies on the use of "Lipschitz quotients" i.e. ratio between output and input distances. A similar method applied to time series (called 'geometrical technique') was presented last year at this workshop [10].

Though different in practice, these methods rely on a common assumption on the continuity of the underlying mapping, and use a geometrical approach based on the data alone. The continuity argument means that if there is a mapping between $\mathbf{x}^{(t)}$ and x_t, then close inputs $\mathbf{x}^{(u)}$ and $\mathbf{x}^{(v)}$ should correspond to close outputs x_y and x_u. Accordingly, as long as the input space is insufficient (i.e. missing delays), close inputs can correspond to arbitrarily distant outputs. Quantifying this is done either by measuring empirical probabilities that two outputs are close given that the corresponding inputs are close (δ-test), or by calculating the ratio between output and input distances (Lipschitz quotients).

It should be noted that these methods are non-parametric. They rely on the data alone, and need not specify a given model (contrary to the ERD method). This can turn out to be a disadvantage since for a given data set, they always select the same set of relevant delays, regardless of the ability of our model to actually implement the underlying mapping. It could very well be that for the model at hand, the estimation would benefit from the inclusion of a *secondary* delay, as shown in the next section and discussed further down. Furthermore, these geometrical techniques require extensive calculations, as they consider all pairs of data. They are thus computationally expensive.

TIME SERIES EXPERIMENTS

This section is devoted to two simple experiments. First we use an artificial problem (the Hénon map), for which a large validation set confirms the results obtained by our ERD method. In the second experiment, we discover interesting long term dependencies on a real time series.

The Hénon map is implemented by the following mapping: $x_t = 1 - 1.4x_{t-1}^2 + 0.3x_{t-2}$. We generate a training set containing 500 data, and a test set of 10000 elements for assessing generalisation abilities. We experiment on non-noisy as well as noisy data, with $\sigma_\varepsilon^2 = 0.1$. Two different models are used: a linear model (obviously ill-suited to this purpose) and a non-parametric kernel smoother. The generalisation estimators are the FPE and LOO respectively.

In order to check whether the delays are wisely chosen, experiments are performed comparing the ERD method and other selection methods (table 1):

1. a forward selection methods using a large validation set (distinct from the test set) of 10000 data;
2. the F_{99}-inclusion, a selection scheme based on the F-statistics [6];
3. the δ-test [14].

As shown on table 1, all forward selection methods outperform the δ-test in the linear case: a linear combination of the first two delays is obviously insufficient to model the mapping. The performance is rather homogeneous among forward selection methods, though the ERD method tends to favour parsimonious models, while keeping good generalisation abilities.

On the non-noisy data, the kernel smoother captures the underlying mapping in all cases. When the training data is noisy, the F-inclusion scheme displays a severe case of curse of dimensionality. The other methods select

Hénon map:		No noise		Noisy	
		Linear	Kernel	Linear	Kernel
Large validation set	Delays	1–7	1–2	1–7	1–3
	MSE	0.376	0.000	0.503	0.214
	Generalisation	0.379	0.000	0.389	0.067
ERD	Delays	1,3–6	1–2	1,3,4	1–3
	MSE	0.376	0.000	0.523	0.214
	Generalisation	0.378	0.000	0.409	0.067
F_{99}-inclusion	Delays	1–6	1–2	1–6,10	1–8,11–13, 16,17,19,20
	MSE	0.376	0.000	0.499	0.032
	Generalisation	0.379	0.000	0.389	0.294
δ-test	Delays	1–2		1–2	
	MSE	0.457	0.000	0.567	0.266
	Gener.	0.455	0.000	0.459	0.097

Table 1: Results on the noisy and non-noisy Hénon map data, for two models: a linear model and a non parametric Kernel smoother. MSE is the Mean Squared (training) Error, generalisation is estimated on 10000 non-noisy data.

one additional delay $t-3$. As we will discuss later, this theoretically unnecessary input leads to an improved prediction accuracy on both the training and generalisation set.

Fraser river data. As an example of real time series processing, we will use a publicly available dataset[2] containing the mean monthly flow of the Fraser River in Hope, British Columbia, from march 1913 to December 1990 [9]. It is a roughly periodic data set containing 946 measurements with maxima every 11 to 13 months. We split the data set so that we have half the data for training and half for testing the prediction abilities of the model. In the following experiments, we use the log values of the data, and estimate the parameters by minimising the Mean Squared Error on the transformed data.

The use of a large validation set is not possible here as is (unfortunately) the case with most real life problem. We will compare the result of the ERD scheme to the results provided by the non-parametric δ-test. According to this test, the embedding space of the time series involves 6 delays.

Note that the ERD method once again outperforms the method based on estimating the embedding dimension. The linear model probes further into the past, and spots relevant delays up to $t-48$, i.e. four times the time span covered by the δ-test. The kernel smoother seems to be experiencing some problems coping with the dimensionality of the data—they could probably be minimised using a variable metric. The neural networks model selects the same amount of delays than the δ-test. However, the selection is targeted towards minimisation of generalisation error, which is reflected in the sizeably smaller test error. Noticeably, the non-linear neural network model, though

[2]available on Statlib at `http://lib.stat.cmu.edu/datasets/`

Fraser river :		Linear	Kernel	Neural net
ERD	Delays	1,2,4–7,10,11, 23,26,35,48	1,2,4,7,11,13	1,2,4,7,11,23
	MSE	0.0529	0.0389	0.0425
	Generalisation	0.0439	0.0547	0.488
δ-test [14]	Delays	1,2,4,7,8,11		
	MSE	0.0680	0.0441	0.452
	Generalisation	0.0609	0.0530	0.627

Table 2: Results for the Fraser river data set, and three different models. MSE is the Mean Squared Error.

using a combination of regularised cost optimisation and OBS pruning [4], does not manage to extract longer-term delay, and is outperformed by the simpler linear model.

OPTIMISING FIR FILTERS

We will now extend the method and apply it to fMRI signal modelling. The fMRI signal measures the hemodynamic response to focal neuronal activation. The data is collected as a 504 steps time-series containing measurements corresponding to the hemodynamic response to a series of periodic baseline and activation periods (7 periods in all). The data is corrupted by a very high level of noise.

Modelling this response as a function of the activation signal is the object of active current research [13]. We extend the above method to optimise the choice of relevant delays when trying to model the response with a FIR filter applied on the excitation signal. Current attempts at doing so use a fixed lag of 7 delays.

The ERD method is simply extended by testing sequentially chosen delays in the excitation signal rather than the time series itself. We applied the method on 5 voxels that were identified as being particularly responsive to the excitation. Out of the 504 measurements, we set the last two periods, or 144 data, aside for testing the generalisation abilities. The first 5 periods, containing 360 points, are used for identifying the relevant delay and the filter coefficients. The FPE is used as a generalisation estimate.

On the 5 fMRI time series studied, we extracted from 1 to 4 delays, ranging from $t-1$ to $t-22$. On voxel number 3 for example, our experiments surprisingly select only $t-1$, but we can see on figure 1 (left panel) that this actually leads to a slight decrease in generalisation error compared to the fixed 7 delay filter. Overall, the results displayed on the left panel of figure 1 suggest that on the extremely noisy data, the method leads to performance that is comparable to the fixed FIR filter, while using less parameters.

On the first voxel, the extraction of relevant delays leads to a noticeable decrease in generalisation error. The right panel of figure 1 plots the response of voxel 1 together with both FIR estimation.

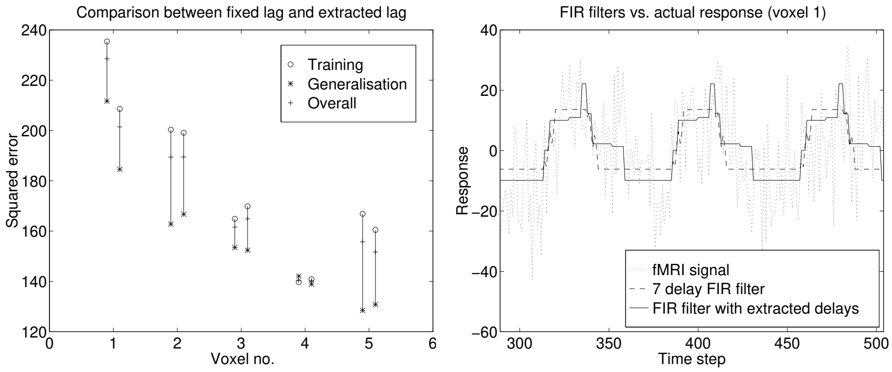

Figure 1: Left: performance comparison. For each voxel the 7 delays FIR filter is on the left, while the FIR filter with extracted delays is on the right. Right: behaviour of the 7 delay filter and the extracted filter on the fMRI time series measured in the first voxel.

DISCUSSION

The FIR example above illustrates the fact that using a parsimonious model, with delays appropriately chosen, is a good way of obtaining good modelling abilities. This can be of great help when facing a problem on which we have no—or little—physical insight. In that context, the ERD is a principled model-dependent approach that has the ability to select the inputs that lead to the best expected prediction error.

It should also be emphasised that it seeks to optimise the actual criteria of interest, i.e. generalisation error. Indeed, at the end of the day we are interested in obtaining the best possible predictions. Reconstructing the dynamics of a time series, as suggested by the methods aimed at estimating the embedding dimension, is only a way of reaching this ultimate goal. On the contrary, the ERD method that we present here tries to optimise the relevant performance criterion directly.

This has an interesting effect: by essence, the ERD method takes into account the fact that modelling is performed on a limited amount of data. On the Hénon map example, this leads to the selection of an additional delay. It has no link to the actual dynamics of the system, but gives a clear decrease in error. Furthermore, when the model is not flexible enough to implement the system mapping, we will probe further into the past, and possibly discover higher-order dependencies that will ease the modelling. This is well illustrated by the two time series examples.

Another aspect of the delay extraction procedure as proposed here is that it relies on the estimation of the generalisation error. It is expected that the more accurate the prediction is, the more relevant the delays selected will be. It should be noted however that we are only interested in finding minima of the generalisation error, so an estimator will be useful as long as it gives the right "trend" in generalisation.

Lastly, let us recall that this method is inspired from the *forward selection* methods in statistical variable selection. A natural extension of this is the use of *backward elimination* steps, in a manner similar to stepwise regression. Similarly, pruning techniques can be used to remove inputs of the model that are potentially harmful with respect to generalisation error.

SUMMARY

We have presented a generalisation-based method for finding the relevant delays in time series modelling. It relates to stepwise variable selection procedures in classical (linear) regression. This 'Extraction of Relevant Delays' method is straightforward to implement and leads to interesting results. When the model is not flexible enough to implement the underlying mapping, it selects additional delays in order to minimise estimated generalisation performance. Noticeably, it outperforms some non-parametric methods for determining the embedding dimension when applied to insufficiently flexible models.

Directions for future work include refinement of the relevance criterion, as well as the extension of this scheme to different problems such as system identification, with more than one temporal inputs.

Acknowledgements. I am very grateful to Jan Larsen for extremely valuable discussions on variable selection. This work was supported by a research fellowship from the Technical University of Denmark.

REFERENCES

[1] H. Akaike, "Fitting autoregressive models for prediction," **Annals of the Institute of Statistical Mathematics**, vol. 21, pp. 243–247, 1969.

[2] T. Cibas, F. Fogelman Soulié, P. Gallinari and S. Raudys, "Variable selection with Optimal Cell Damage," in **Proceedings of ICANN'94**, 1994, pp. 727–730.

[3] C. Goutte, "On the use of a pruning prior for neural networks," in **Neural Networks for Signal Processing VI – Proceedings of the 1996 IEEE Workshop**, 1996, pp. 52–61.

[4] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in S. Hanson, J. Cowan and C. Giles, eds., **Advances in Neural Information Processing Systems**, Morgan Kaufmann, 1993, vol. 5 of **NIPS**, pp. 164–171.

[5] X. He and H. Asada, "A new method for identifying orders of input-output models for nonlinear dynamic systems," in **American Conference on Control**, San Francisco, California, 1993.

[6] R. R. Hocking, "The analysis and selection of variables in linear regression," **Biometrics**, vol. 32, pp. 1–49, March 1976.

[7] J. Larsen and L. K. Hansen, "Generalized performance of regularized neural networks models," in **Neural Networks for Signal Processing IV – Proceedings of the 1994 IEEE Workshop**, 1994, pp. 42–51.

[8] J. Larsen and L. K. Hansen, "Empirical generalization assessment of neural network models," in **Neural Networks for Signal Processing V – Proceedings of the 1995 IEEE Workshop**, 1995, pp. 42–51.

[9] A. I. McLeod, "Diagnostic checking of periodic autoregression models with application," **Journal of Time Series Analysis**, vol. 15, no. 2, pp. 221–233, 1994.

[10] C. Molina, N. Sampson, W. J. Fitzgerald and M. Niranjan, "Geometrical techniques for finding the embedding dimension of time series," in **Neural Networks for Signal Processing VI – Proceedings of the 1996 IEEE Workshop**, 1996, pp. 161–169.

[11] J. Moody, "Note on generalization, regularization and architecture selection in nonlinear learning systems," in **Proceedings of the first IEEE Workshop on Neural Networks for Signal Processing**, 1991, pp. 1–10.

[12] N. Murata, S. Yoshizawa and S. Amari, "Network Information Criterion—determining the number of hidden units for an artificial neural network model," **IEEE Transactions on Neural Networks**, vol. 5, no. 6, pp. 865–872, 1994.

[13] F. Å. Nielsen, L. K. Hansen, P. Toft, C. Goutte, N. Lange, S. C. Strother, N. Mørch, C. Svarer, R. Savoy, B. Rosen, E. Rostrup and P. Born, "Comparison of two convolution models for fMRI time series," in **3rd Intl. Conference on Functional Mapping of the Human Brain**, 1997.

[14] H. Pi and C. Peterson, "Finding the embedding dimension and variable dependences in time series," **Neural Computation**, vol. 6, no. 3, pp. 509–520, May 1994.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, **Numerical Recipes in C**, Cambridge University Press, 2nd edn., 1992.

[16] R. Savit and M. Green, "Time series and dependent variables," **Physica D**, vol. 50, no. 1, pp. 95–116, 1991.

[17] G. Toussaint, "Bibliography on estimation of misclassification," **IEEE Transactions on Information Theory**, vol. 20, no. 4, pp. 472–479, July 1974.